

Poisson PCA for matrix count data

J. Virta¹ A. Artemiou²

¹University of Turku

²University of Limassol

EcoSta 2023, 1st of August

Table of Contents

- 1 Normal factor model
- 2 Poisson factor model
- 3 Data example
- 4 Simulation study
- 5 Closing remarks

- Let X_1, \dots, X_n be a sample of $p_1 \times p_2$ matrices, representing, for example, abundances of n species in combinations of p_1 areas and p_2 time periods.

Abundance matrix for the i th species

$$\begin{array}{c} \text{Area}_1 \\ \vdots \\ \text{Area}_{p_1} \end{array} \begin{pmatrix} \text{Time}_1 & \cdots & \text{Time}_{p_2} \\ X_{i,11} & \cdots & X_{i,1p_2} \\ \vdots & \ddots & \vdots \\ X_{i,p_11} & \cdots & X_{i,p_1p_2} \end{pmatrix}$$

- For large dimensions p_1, p_2 , the first step of data analysis is often *dimension reduction*.

- One of the simplest models for the dimension reduction of matrix data is the normal factor model,

$$X = \mu + U_1 Z U_2^T + \varepsilon,$$

where

- $Z \sim \mathcal{N}_{d_1 \times d_2}(0, \tau \Lambda_1, \tau \Lambda_2)$ is a latent matrix with independent elements.
- The scales Λ_1, Λ_2 are diagonal matrices.
- The loadings U_1, U_2 have orthonormal columns.
- The error $\varepsilon \sim \mathcal{N}_{p_1 \times p_2}(0, \sigma I_{p_1}, \sigma I_{p_2})$ is independent of Z .

- The normal model is not suitable for abundance data which takes non-negative integer values.
- To later extend the normal model to such count-valued data, we prefer to write it in conditional form:

$$\begin{cases} Z \sim \mathcal{N}_{d_1 \times d_2}(0, \tau \Lambda_1, \tau \Lambda_2) \\ X | Z \sim \mathcal{N}_{p_1 \times p_2}(\mu + U_1 Z U_2^\top, \sigma I_{p_1}, \sigma I_{p_2}) \end{cases}$$

- Thus, the model is a *Normal-Normal mixture*.

- Assuming that the latent dimensions d_1, d_2 are known, the model parameters can be estimated from the sample versions of the following moments

$$\mathbb{E}(X),$$

$$\mathbb{E}[\{X - \mathbb{E}(X)\}\{X - \mathbb{E}(X)\}^\top],$$

$$\mathbb{E}[\{X - \mathbb{E}(X)\}^\top \{X - \mathbb{E}(X)\}].$$

- The latent variables are estimated as

$$U_1^\top (X - \mu) U_2 = Z + \varepsilon_0.$$

Table of Contents

- 1 Normal factor model
- 2 Poisson factor model**
- 3 Data example
- 4 Simulation study
- 5 Closing remarks

Poisson-Normal mixture

$$\begin{cases} Z \sim \mathcal{N}_{d_1 \times d_2}(0, \tau \Lambda_1, \tau \Lambda_2) \\ X | Z \sim \text{Poisson}_{p_1 \times p_2} \{ \exp(\mu + U_1 Z U_2^T) \}, \end{cases}$$

where $\text{Poisson}_{p_1 \times p_2}(M)$ has independent Poisson-elements with mean matrix M .

- The latent variables are normal (continuous)
- They influence the observed counts through an exponential link function.

Vector Poisson-Normal mixture

- If $p_2 = 1$, we obtain a vector Poisson-Normal mixture that was originally proposed by [Aitchison and Ho, 1989].
- They called it the Poisson-Lognormal model.
- The model was later studied by [Hall et al., 2011, Kenney et al., 2021].

- The method of moments yields closed form solutions for the model parameters.

$$\theta := (\mu, U_1, \Lambda_1, U_2, \Lambda_2, \tau).$$

- For example, the left loadings U_1 and scale Λ_1 can be estimated from the eigendecomposition of the matrix S_1 defined as,

$$s_{1,jk} := \frac{1}{p_2} \sum_{\ell=1}^{p_2} \log \left\{ \frac{E(x_{j\ell}x_{k\ell})}{E(x_{j\ell})E(x_{k\ell})} \right\},$$
$$s_{1,jj} := \frac{1}{p_2} \sum_{\ell=1}^{p_2} \log \left[\frac{E\{x_{j\ell}(x_{j\ell} - 1)\}}{\{E(x_{j\ell})\}^2} \right],$$

- Standard limiting theory shows that the resulting sample estimator θ_n converges to a limiting normal distribution at root- n rate.
- The limiting covariance matrix has a particularly cumbersome form.

Latent component estimation

- We estimate the latent matrix Z as the mode of the conditional distribution $Z \mid X$.
- The resulting distribution is not of standard form but its density is log-concave and admits a unique mode.
- Standard gradient descent can be used for the estimation.

- We estimate the latent dimensions d_1, d_2 using *predictor augmentation* [Luo and Li, 2020].
 - The dimension d_1 equals the rank of the matrix S_1 .
 - We augment the observed data as

$$X_i^* = \begin{pmatrix} X_i \\ R_i \end{pmatrix},$$

where R_i have iid Poisson(1)-elements.

- By comparing the sample estimates S_{n1} and S_{n1}^* it is possible to identify where the d_1 -dimension signal “ends”.
- [Luo and Li, 2020] prove consistency of this procedure but their proof does not apply to discrete data.

Augmentation curve

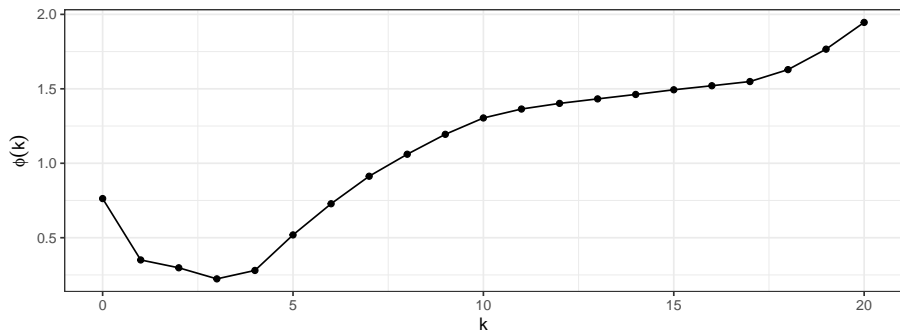


Figure: The minimum value of the augmentation curve is achieved at $d_1 = 3$.

Table of Contents

- 1 Normal factor model
- 2 Poisson factor model
- 3 Data example**
- 4 Simulation study
- 5 Closing remarks

- Abundance data available at <https://github.com/rfrelat/Multivariate2D3D>.
- The data consists of abundances of a total of $n = 65$ fish species
 - in seven areas (RA 1 – RA 7), $p_1 = 7$,
 - during 6 time periods (1985 – 1989, ..., 2005 – 2009, 2010 – 2015), $p_2 = 6$.
- [Frelat et al., 2017] identified six biologically meaningful clusters (*Southern*, *Northern*, *NW Increasing*, *SE Increasing*, *Increasing* and *Decreasing*) in the data.

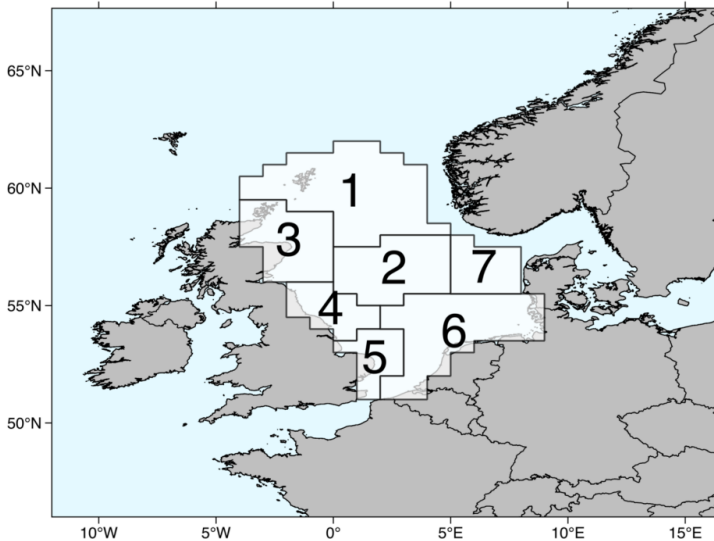


Figure: Image from [Frelat et al., 2017].

Dimension estimation

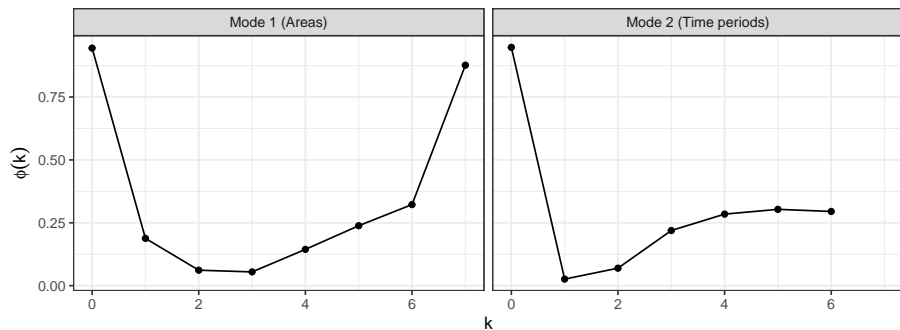


Figure: The dimensions are estimated as $d_1 = 3$ and $d_2 = 1$.

Latent components

- We estimate a total of three latent components, $z_{i,11}$, $z_{i,21}$ and $z_{i,31}$.
- The first of these turns out to measure the overall abundances of the species.
- We plot the remaining two together with the column loadings into a *biplot*.

Biplot

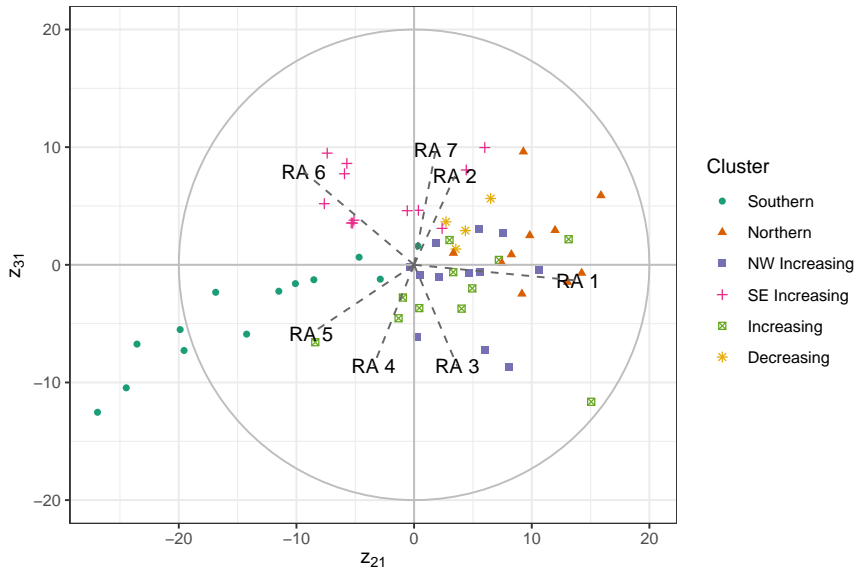


Table of Contents

- 1 Normal factor model
- 2 Poisson factor model
- 3 Data example
- 4 Simulation study**
- 5 Closing remarks

- We simulated samples of 4×3 matrices from the Poisson factor model and estimated its parameters using three different methods:
 - Our proposal.
 - Vectorizing and method of moments [Aitchison and Ho, 1989].
 - Vectorizing and MLE with variational inference [Hall et al., 2011].
- Average errors over 1000 replicates for various sample sizes and covariance structures are shown on the following slide.

Efficiency plot

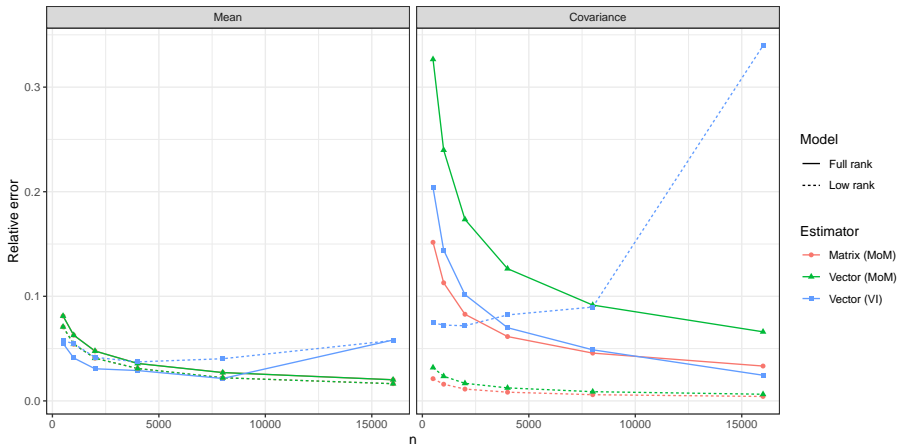





Figure: Average estimation errors of the three methods.




Table of Contents

- 1 Normal factor model
- 2 Poisson factor model
- 3 Data example
- 4 Simulation study
- 5 Closing remarks

Some remarks

- The presentation is based on [Virta and Artemiou, 2023].
- Many count-valued data exhibit sparsity \rightarrow zero-inflated variant?
- Bernoulli-Normal mixture for binary matrix data?

-  Aitchison, J. and Ho, C. (1989).
The multivariate Poisson-log normal distribution.
Biometrika, 76(4):643–653.
-  Frelat, R., Lindegren, M., Denker, T. S., Floeter, J., Fock, H. O., Sguotti, C., Stähler, M., Otto, S. A., and Möllmann, C. (2017).
Community ecology in 3D: Tensor decomposition reveals spatio-temporal dynamics of large ecological communities.
PLoS one, 12(11):e0188205.
-  Hall, P., Ormerod, J. T., and Wand, M. P. (2011).
Theory of Gaussian variational approximation for a Poisson mixed model.
Statistica Sinica, pages 369–389.

-  Kenney, T., Gu, H., and Huang, T. (2021).
Poisson PCA: Poisson measurement error corrected PCA, with application to microbiome data.
Biometrics.
-  Luo, W. and Li, B. (2020).
On order determination by predictor augmentation.
Biometrika.
-  Virta, J. and Artemiou, A. (2023).
Poisson PCA for matrix count data.
Pattern Recognition, 138:109401.